

MiniBooNE Computing Infrastructure

Chris Green
Fermilab

Experiment specifics

- Purpose
 - Measure neutrino and anti-neutrino interactions from the Booster.
- Number of users - **~100 cumulative total**
 - Remote - **~50**
 - Using Fermilab facilities **~50**

Experiment schedule

	Pre-2009	2009	2010	2011	2012	2013	2014
Planning	X						
Construction	X						
Commissioning	X						
Data taking	X	X	X	X	X		
Data analysis	X	X	X	X	X		

Data

- How many events/year?
 - Test beam: 0
 - Pedestal and calibration: 77M
 - Normal data: 33M
 - Normal data after quality filtering?: 30M
- How large is each event?
 - Zero suppressed: 13 KBytes
 - Non zero-suppressed: N/A
 - Reconstructed: 13 KBytes
 - Simulated: 100 KBytes
 - Data summary? N/A

Central FNAL systems

- CPU used (see table)
- Storage used (see table)
- Uses:
 - AFS
 - Reconstruction and data filtering
 - Calibration and alignment
 - MC Generation
 - User data analysis

Data flow

	Pre-2009	2009	2010	2011	2012	2013	2014
Raw Data, TB	~5	1	1	1	1	0	0
Processed Data, TB	~5	1	1	1	1	0	0
User data, TB	0	0	0	0	0	0	0
Simulated data, TB	~40	1	0	0	1	0	0

Please enter incremental quantities

CPU needs

	Pre-2009 (/y)	2009	2010	2011	2012	2013	2014
Running	3	3	3	3	3	0	0
Reconstruction	50	50	0	0	100	0	0
Calibration	.5	.5	.5	.5	.5	0	0
Skimming	.5	.5	.5	.5	.5	0	0
Analysis	100	50	10	0	10	0	0
Simulation	100	0	0	0	40	0	0

Please use CPU-years on a current machine

e.g. # events * time per event in sec * 3×10^7 * reprocessing factor

Operating systems

- What OS is used?
 - FSL 4.
- Do all collaborators have to use the same one?
 - Must be binary-compatible if they wish to use the software.

Data storage and tracking

- How do you catalog data?
 - Home-grown allocation and link management system.
- How do you provide remote access to data?
 - All use of primary data takes place at FNAL.

Remote systems

- How many remote institutions provide resources for your users/collaboration
 - Many institutions provide desktop computing based at FNAL integrated into cycle-stealing Condor system (>50 machines at all times).
 - Colorado, LANL have provided batch computing.
- Do they have special systems for you or shared?
 - Shared.
- What is done at remote institutions?
 - Some MC generation, some reprocessing.
 - User analysis of secondary data sets (ntuples).

Data distribution to remote sites

- Where are data distributed
 - From / to Colorado, LANL.
- What kind of data
 - Auxiliary input data for MC, MC output.
 - Input data, reprocessed output data.
- How much data
 - ~0.5TB.
- How fast does it need to move
 - On demand.
- What method is used
 - GridFtp.

Grid

- Do you use the Grid? **Yes**
- Do you use Grid tools such as Gridftp? **Yes**
- Do you use Glidein or some other tool? **No**
- Do you use the FNAL Grid exclusively or do you use more general grid resources?
 - **Almost exclusively FNAL.**

Databases

- Technology used – PostgreSQL
- Size – 14GB
- Access rate – av. few times/h, large fluctuations.
- Are they replicated remotely? No.
- What is stored? Calibrations, run quality info, run summary info, other meta-information.

Conditions

- How are conditions and calibrations stored?
 - PostGresQL DB.
- How are they accessed? Custom FORTRAN / C++ / shell / perl wrappers around PostGres API.

Code management

- Code repository
 - CVS
- Build system
 - SoftRelTools
 - CMT (for Nuance)

Standard packages

- What standard packages are used:
 - HBOOK *et al*
 - GEANT3
 - GEANT4
 - ROOT
 - NUANCE
 - RCP
 - ENCP

What worked really well?

- Local cycle-stealing Condor on desktops for user analysis.
- Grid-based process – huge step forward from old farm in terms of throughput.
- AFS for user home systems, software distribution, etc, married quite well with large number of user desktops.

What would you not do again?

- Home grown data catalog made somewhat redundant by bluearc.
- AFS integrated into near line processing.
- Near line writing straight to Enstore (see Art's talk)!

Other nuggets

- Do not underestimate the amount of work your users will want to do. Interactive use of one desktop between 1-5 people may not cut it.
- Users feel far more strongly than you might think about the difference between interactive and batch. Your task is to make that difference as small as possible one way or another.
- Do not underestimate the work involved in keeping your software operating on modern OS.